ARTICLES

D-Lib Magazine
November/December 2009

Volume 15 Number 11/12

ISSN 1082-9873

Beyond 1923

Characteristics of Potentially In-copyright Print Books in Library Collections

Brian Lavoie
Lorcan Dempsey
OCLC Online Computer Library Center
{lavoie, dempseyl}@oclc.org

Introduction

Issues of copyright and permissible use have swirled around efforts to digitize print book collections. Sharp debate has ensued over the circumstances in which creating a digital surrogate and making it accessible online runs afoul of copyright protections, and what remedies might be appropriate to compensate rights holders. Some digitization efforts, such as the Open Content Alliance, have restricted themselves to public domain materials; Google Books, on the other hand, has sought to reach agreement with copyright holders represented by the Authors Guild and the Association of American Publishers. A proposed class action settlement,[1] announced in October 2008, would create a Book Rights Registry responsible for administering and adjudicating the process of locating and compensating rights holders impacted by Google's digitization activities.

The Google book settlement provoked spirited discussion of its potential ramifications, mimicking the commotion that followed the announcement of the original Google Print for Libraries (later re-named Google Books) project in December 2004. Using data from the WorldCat bibliographic database,[2] OCLC Research published an article in 2005 aimed at illuminating issues surrounding Google's plan to digitize the print book collections of five major research libraries. The present article is motivated by a similar purpose: to provide empirical context for the many discussions surrounding the digitization of in-copyright print books. The settlement has raised challenging questions regarding permissible use of print book titles published after 1923; many of these titles may eventually form a significant part of the Google book database should it come to pass.

Discussions of Google Books and other digitization efforts tend to treat in-copyright print books as an amorphous collection, with little elaboration or detail on what this important collection of materials actually looks like. How many titles are involved? What is the distribution of their

publication dates? What general observations can be made about their content? This article examines these and other questions in regard to the collection of US-published print books represented in WorldCat. Many of these questions were posed to the authors in private inquiries; these inquiries, along with the keen interest in digitization that continues to spark debate on blogs and listservs, suggested that a general publication addressing the characteristics of in-copyright print books could provide helpful context for ongoing discussions.

The focus of this article is on print book titles that are either in-copyright or potentially in-copyright. Determining copyright status is, however, problematic. The nuances of US copyright law are quite complicated, but a useful simplification organizes print books into three categories of copyright status based on date of publication. Broadly speaking, works published before 1923 are considered in the public domain, and therefore unencumbered by copyright restrictions. The copyright status of books published between 1923 and 1963, however, is murkier. Under US copyright law, works published during this period with a copyright notice remain in copyright for 95 years after publication – *if their copyright was renewed*. If copyright was allowed to lapse, the work reverts to the public domain. Finally, books published after 1963 are, by and large, still in copyright.

In addition to copyright status, the question of *orphan works* has received much attention in regard to digitization activities. The United States Copyright Office defines an orphan work as "the situation where the owner of a copyrighted work cannot be identified and located by someone who wishes to make use of the work in a manner that requires permission of the copyright owner."[3] While it is important to bear in mind that any in-copyright book can be an "orphan", in practice the prevalence of orphan works is likely to be skewed toward older, rather than recently published, materials.

The analysis that follows examines the characteristics of US-published print books, with an emphasis on books that are likely in copyright according to US copyright law.[4] As with our earlier article, the analysis is based on data from the WorldCat database, which represents the aggregated collections of more than 70,000 libraries worldwide. The analysis focuses on three areas: the WorldCat aggregate collection of US-published print books; the subset of this collection published during or after 1923 – i.e. those potentially associated with copyright and/or orphan works issues; and the combined print book collection of three academic research library participants in Google Books – again, with an emphasis on materials that are potentially in copyright.

Characteristics of the aggregate US-published print book collection in WorldCat

As of April 2009, the WorldCat bibliographic database contained about 135.3 million bibliographic records representing information resources of all descriptions. Of these, 104.1 million represented books, and of these, 84.8 million were print books.[5] Finally, of these, 15.5 million were print books published in the US, and therefore presumably covered by US copyright law (Figure 1). It is important to keep in mind that these counts do not represent "physical objects" – i.e., copies of books on the shelf – but rather, distinct imprints or *manifestations*.[6] For example, the Short Books, Limited publication of the book *Walking Ollie*, published in London in 2006, is a distinct print book manifestation; the version of the same book published by Perigee Books (New York, 2008) is another distinct manifestation. There are likely hundreds if not thousands of physical copies of these two manifestations worldwide in the thousands of institutional print book collections represented in WorldCat. However, each manifestation would only be counted once in WorldCat.

For the remainder of this article, we will refer to these 15.5 million print book manifestations as the aggregate US-published print book collection in WorldCat.

Figure 1: US-published Print Books in WorldCat: April 2009

The aggregate US-published print book collection is a resource that is collectively managed by many institutions; the 15.5 million distinct manifestations in the aggregate collection inflate to 656.8 million total holdings in library collections around the world, where a "holding" simply means that a particular library collection contains at least one copy of a particular print book manifestation. In fact, the 15.5 million US-published print book manifestations – constituting only 11 percent of the materials represented in the WorldCat database – account for 46 percent, or nearly half, of the more than 1.4 billion total holdings attached to materials represented in WorldCat. To some extent, this result is predictable: North American library collections are especially well-represented in WorldCat, and one would expect North American libraries to collect US publishing output heavily. Certainly these libraries also collect much of the rest of the world's published output as well, but their combined collections yield a strong concentration of US-published print books.

Probing a little deeper into the total holdings of the aggregate US-published print book collection reveals that a wide range of institution types is represented within the group of libraries that collectively hold this resource. Table 1 lists the percentage of total holdings attributable to different institution types for the aggregate US-published print book collection.

Table 1: Total holdings of the aggregate US-published print book collection, by institution type

Type	Number	Percentage
Academic	363,542,724	56 percent
Public	219,920,744	33 percent
Special	20,898,191	3 percent
School	19,147,728	3 percent
Other Government	11,816,402	2 percent

State & National	8,660,213	1 percent
Other	5,816,423	1 percent
Type unknown	7,021,683	1 percent

More than half of the holdings attached to the 15.5 million US-published print book manifestations belong to academic institutions, while a third belong to public libraries, and the rest to a variety of other institution types. This suggests that among collecting institutions, academic libraries possess an especially considerable stake in issues impacting accessibility, use, and preservation of US-published print books, if for no other reason than by virtue of the comparatively large investment they have made in collecting them, and the consequently large presence these materials have in their collections.

The age (i.e., publication date) of the titles in the US-published aggregate print book collection is not distributed evenly over time, but instead is skewed toward newer materials. Figure 2 shows the distribution of US-published print book manifestations in WorldCat, by publication year, for the period 1900-2008.

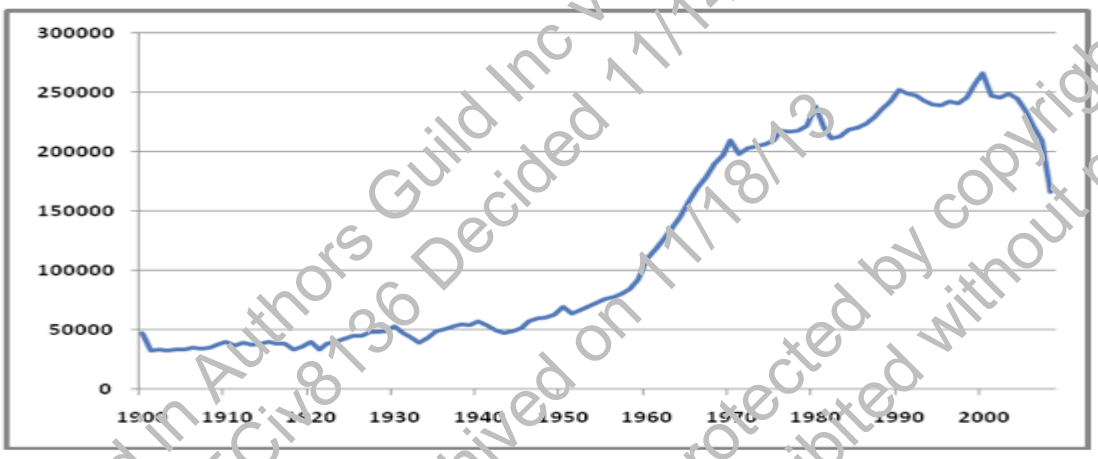

Figure 2: US-published print book manifestations, by publication date (1900-2008)*

* Note: the drop in manifestation totals after 2000 is not a consequence of decreased publishing output or collecting activity in those years, but instead represents "cataloging lag", the elapsed time between a book's date of publication and the date that a bibliographic record for the book is entered into the WorldCat database. Cataloging lag can be divided into two components: 1) the *acquisition lag*, which is the time elapsed between the date of publication and the time it is acquired by a library; and 2) the *processing lag*, which is the time between the date of acquisition and the date a record for the book is entered into WorldCat.

Approximately half of the manifestations in the aggregate US-published print book collection in WorldCat were published after 1977; two-thirds were published after 1964, and three-quarters after 1951. This would suggest that the number of US-published print book manifestations in library collections that are "free and clear" in terms of copyright restrictions is comparatively small, while the fraction that is likely in copyright is comparatively large. Table 2 sharpens this point by organizing US-published print book manifestations in WorldCat by the three broad time frames relevant to assessing copyright status.

Table 2: Distribution of US-published print book manifestations in WorldCat, by major US copyright period

Period	Number	Percentage
Pre-1923:	2,227,048	14 percent
1923-1963:	2,596,114	17 percent
Post-1963:	9,991,301	65 percent
Unknown/questionable date:	667,814	4 percent

The percentages reported in Table 2 indicate that about 14 percent of the US-published aggregate print book collection was published before 1923, and therefore is, with reasonable certainty, in the public domain according to US copyright law. A further 17 percent were published between 1923 and 1963; for these, copyright status cannot be ascertained without investigating each individual title. Some portion of these materials will be in the public domain – in particular, those whose copyright was not renewed. The rest will still be under copyright. Recent statistics from the HathiTrust indicate that about 60 percent of candidate materials for digitization published between 1923 and 1963 reverted to the public domain, either because copyright was not renewed, the book was published without a copyright notice, or for other reasons.[7] Applying this fraction to the US-published aggregate print book collection in WorldCat suggests that approximately 1.6 million manifestations are public domain, while the remaining 1 million are still in copyright.

The HathiTrust result is based on academic library holdings, while the aggregate print book collection in WorldCat represents the holdings of a variety of institution types (although as Table 1 indicates, academic libraries hold the largest portion). A more general, but much earlier study by the US Copyright Office in 1960 found that only 7 percent of books registered for copyright in 1931-32 had had their copyright renewed within the prescribed 28 year period after initial registration. The remainder of the books would have reverted to the public domain.[8] Both the HathiTrust and Copyright Office results suggest that of the print books published between 1923 and 1963, a majority – and perhaps a substantial majority – are likely to be in the public domain.

Finally, about two-thirds of the US-published aggregate print book collection represents books that were published after 1963, and therefore are almost certainly still in copyright.

As noted at the beginning of this article, an orphan work is any work under copyright where the rights holder cannot be identified or located. In light of the US-published aggregate print book collection in WorldCat, this issue is relevant for the 12.6 million print book manifestations that could *potentially* still be in copyright – the sum of the totals for the two periods 1923-1963 and post-1963 in Table 2, or 82 percent of the collection. Of course, in practice the fraction that end up as orphan works will be considerably smaller; the 12.6 million manifestations is appropriately interpreted as the pool of materials from which orphan works could potentially emerge. However, even if but a fraction of these manifestations end up as orphan works – say as little as 10 percent – that would still account for over 1 million manifestations where administration of copyright permissions is inhibited by a lack of a clear rights holder. And this is not the full extent of the orphan works problem: orphan works can emerge from books published outside the US as well.

Characteristics of potentially "in copyright" US-published print books

In this section, the characteristics of the 12.6 million print book manifestations published during or after 1923 – that is, the pool of print book manifestations in library collections that are potentially in copyright – are examined as an aggregate collection. It is these materials that pose potential copyright-related difficulties in the context of print book digitalization efforts such as Google Books and others. A number of aspects of these materials are examined in this section, including authors,

subject, and audience level, with a special focus on nonfiction materials.

The aggregate collection of potentially in-copyright, US-published, print book manifestations in WorldCat is associated with about 3.7 million unique authors (individuals)[9] who served some creative role in regard to one or more of the print book manifestations in the collection, and therefore could potentially hold some form of copyright authority over future use of the book. Table 3 reports the authors associated with the most print book manifestations in WorldCat that are potentially still in copyright.

Table 3: Authors associated with the most potentially in-copyright print book manifestations

Author	Number of Print Book Manifestations
William Shakespeare	6,226
Carole Marsh	3,295
Mark Twain	2,979
Jack Rudman	2,554
Charles Dickens	2,359
Ronald Vern Jackson	2,265
Harold Bloom	2,234
Agatha Christie	2,061
Robert Louis Stevenson	1,946
Joy Cowley	1,877

Note that this list is not intended to imply that these are the most "important" authors (although some clearly are!); only that these are the most frequently appearing authors in the collection under study. A perhaps surprising feature of this list is that many of the authors are those whose work one might suppose had long since passed into the public domain. Authors whose works have become time-honored classics will be published and re-published frequently over time; intellectual property rights will likely attach to the various manifestations of these works, in the form of new introductions, illustrations, and even updates, revisions, or new representations of the "main body" of the work itself that constitute sufficient grounds to claim "new material". The names on the list in Table 3 serve as a useful reminder that digitization efforts may encounter copyright issues even with works originally published centuries ago; recently published manifestations of Shakespeare's *Macbeth*, or Dickens' *A Christmas Carol*, for example, are likely not in the public domain.

Of the 12.6 million print book manifestations published during or after 1923, the overwhelming majority – 92 percent, or 11.6 million – were nonfiction; only 8 percent, or about 1 million, were fiction. For the remainder of this section, statistics reported pertain to *nonfiction* print book manifestations only.

Table 4 presents a breakdown of the post-1923 non-fiction print book manifestations by subject.[10]

Table 4: Subject breakdown, nonfiction print books

History and auxiliary sciences	8 percent
Engineering and technology	7 percent
Business and economics	7 percent
Language, linguistics, and literature	6 percent
Philosophy and religion	5 percent
Health and medicine	5 percent
Art and architecture	3 percent
Law	3 percent
Sociology	3 percent
Education	3 percent
Other	15 percent
Unknown	35 percent

The methods used in this study were able to assign subject categories to about two-thirds of the nonfiction print book manifestations. The distribution of the categorized manifestations over subject was fairly even, with no category exceeding 8 percent of the total. History is the most populous category, followed by Engineering & Technology and Business & Economics. These categories account for nearly a quarter of the nonfiction manifestations.

In addition to subject, it is also possible to make some observations about the "audience level" of the nonfiction print book manifestations. Audience level is an inference about the nature of the content of a book. While audience level cannot be determined directly, it is possible to make some useful inferences based on the collecting decisions of libraries. For example, if a particular book is held primarily by academic libraries, we can infer that it is likely intended for a scholarly or research-oriented audience. If a book is held chiefly by primary or secondary education institutions, it is likely intended for a juvenile audience. Examining the library holdings attached to a particular print book manifestation in WorldCat, it is possible to calculate a metric whose value ranges from zero to one, where values closer to zero indicate the book is held mainly by primary or secondary education institutions, and values closer to one would indicate the item is held mainly by academic libraries.[11] For simplicity, we divide this continuum up into three categories:

- *Audience level from 0 to 0.33*: "Schooler" (primarily intended for a juvenile audience)
- *Audience level from 0.33 to 0.67*: "General" (primarily intended for a non-specialist readership)
- *Audience level from 0.67 to 1.0*: "Scholar" (primarily intended for an academic or specialist readership)

It should be noted that the audience level metric is only an estimate, and may not be accurate for any particular book. However, as a means of making some general observations about the broad contours of an aggregate collection of books, it does provide some useful insights.

Application of the audience level metric to the collection of US-published nonfiction print books published during or after 1923 is limited to books whose records were entered into WorldCat prior

to January 2007.__ This restriction helps ensure that the book's record has resided in WorldCat long enough to accumulate a sufficiently representative collection of holdings. This reduces the pool of print book titles to about 8 million. Table 5 reports the audience level calculations for these materials.

Table 5: Audience level breakdown, US-published nonfiction print books published during or after 1923*

Schooler:	4 percent
General:	42 percent
Scholar:	54 percent

* For manifestations where it was possible to calculate an audience level (92 percent of total). Most of the manifestations without an audience level were materials held by a type of library not considered in the audience level calculations.

The vast majority of the nonfiction print book manifestations (96 percent) were intended for general readership or higher; the fraction intended for juvenile audiences was quite small (4 percent). The fraction devoted to a scholarly or specialist audience constituted the largest fraction of the titles. This finding complements the results in Table 1, which indicate that the majority of US-published print book holdings in library collections are attributable to academic institutions.

More nuanced results are obtained by examining audience level calculations for several of the subject categories identified in Table 4. Table 6 reports the audience level breakdown for the top six subject categories in Table 4. As with Table 5, the results reported in Table 6 are limited to print books entered into WorldCat prior to 2007.

Table 6: Audience level breakdown, US-published nonfiction print books published during or after 1923, by subject*

Subject	Schooler	General	Scholar
History	6 percent	58 percent	36 percent
Engineering/Technology	3 percent	45 percent	51 percent
Business/Economics	1 percent	34 percent	66 percent
Language/Linguistics/Literature	8 percent	40 percent	52 percent
Philosophy/Religion	3 percent	52 percent	45 percent
Health/Medicine	2 percent	31 percent	67 percent

* For manifestations where it was possible to calculate an audience level. Fraction of manifestations with an audience level in each subject category: History (94 percent); Engineering/Technology (90 percent); Business/Economics (92 percent); Language/Linguistics/Literature (99 percent); Philosophy/Religion (97 percent); Health/Medicine (95 percent).

According to the results in Table 6, all subject categories had relatively small fractions of materials intended primarily for juvenile audiences; Language/Linguistics/Literature exhibited the highest fraction with 8 percent. History was the subject category with the highest proportion of materials

aimed at a general or non-specialist audience (58 percent); Philosophy/Religion was the only other category where the proportion of General materials exceeded the proportion of Scholar materials. Health/Medicine was the most "scholarly" subject category, with two-thirds of print books aimed at a research or specialist audience; Business/Economics was slightly behind Health/Medicine at 66 percent. Taken together, the results in Table 6 indicate that digitization efforts aimed at subject categories like Health/Medicine and Business/Economics would most likely confer the most benefits on a scholarly or specialist audience; in contrast, digitization aimed at History or Philosophy/Religion would perhaps be of greater benefit to non-specialist readers.

The preceding analysis has focused on the accumulated US non-fiction publishing output from 1923 to the present. To gain a different perspective on the characteristics of these materials, print books published in 1923 and print books published in 2000 were identified and compared as a means of examining changes in the characteristics of nonfiction print books in WorldCat over time. About 40,000 US-published nonfiction print book manifestations published in 1923 reside in WorldCat, compared to about 235,000 published in 2000. Table 7 reports the subject category breakdown for these two collections of print books.

Table 7: Subject breakdown, US-published nonfiction print books (1923 and 2000)

1923:	
Language, Linguistics and Literature:	10 percent
History and Auxiliary Sciences:	9 percent
Philosophy and Religion:	6 percent
Business and Economics:	5 percent
Engineering and Technology:	4 percent
Health and Medicine:	3 percent
Other:	18 percent
Unknown:	44 percent
2000	
History and Auxiliary Sciences:	11 percent
Business and Economics:	7 percent
Engineering and Technology:	7 percent
Language, Linguistics and Literature:	6 percent
Philosophy and Religion:	6 percent
Health and Medicine:	6 percent
Law:	4 percent
Art & Architecture:	4 percent
Sociology:	4 percent

Education:	4 percent
Computer Science:	3 percent
Other:	14 percent
Unknown:	24 percent

Key differences between print books published in 1923 and those published in 2000 include a much larger proportion devoted to Language/Linguistics/Literature in the earlier period, while a larger proportion was captured by Engineering/Technology and Business/Economics in the later period. And of course, Computer Science as a subject category only makes an appearance in the latter year. The results in Table 7 suggest that the so-called STM (science, technology, medicine) subject categories account for a significantly higher proportion of the print book manifestations in 2000 than in 1923. For example, Engineering/Technology and Health/Medicine together account for 13 percent of the manifestations in 2000, compared to 7 percent in 1923. In contrast, the humanities are slightly more predominant in 1923 compared to 2000: History, Language/Linguistics/Literature, and Philosophy/Religion account for 25 percent of the titles in 1923, compared to 23 percent in 2000. However, these findings should be considered preliminary results; more work needs to be done to verify and sharpen their implications. A large percentage of print books published in 1923 were not categorizable by subject according to the methods used in this study. Allocation of these uncategorized manifestations to the various subject categories could shift the proportions significantly. This consideration also impacts the results for 2000, but to a lesser extent: in this case, only about a quarter were not categorized.

Another factor impacting interpretation of the findings in Table 7 is the question of whether the distribution of manifestations across subject categories is influenced more by the scope of publishing output in each year, or by the collecting and retention decisions of libraries. It is not possible to resolve this question from the data in Table 7, but it is an important issue in understanding the dynamic character of the collection of US-published print books in WorldCat.

Table 8 reports the audience level breakdown for US-published nonfiction print books published in 1923 and 2000.

Table 8: Audience level breakdown, US-published nonfiction print books (1923 and 2000)*

	Schooler	General	Scholar
1923	1 percent	22 percent	77 percent
2000	8 percent	54 percent	38 percent

* For manifestations where it was possible to calculate an audience level. Fraction of manifestations published in 1923 with audience level calculation: 92 percent; in 2000: 92 percent.

The results in Table 8 suggest a significant shift between 1923 and 2000 in the distribution of print books across juvenile, non-specialist, and specialist audiences. In 1923, more than three-quarters of the print books are aimed at a specialist, "scholarly" audience; in contrast, less than 40 percent of the print books published in 2000 fall into the Scholar category, while more than half fall within the purview of a general readership. As with the differences in the distribution across subject categories over time, it is difficult to determine whether this difference in audience level proportions between 1923 and 2000 is a result of shifts in publishing trends, or shifts in collecting and retention decisions by libraries.

The potential influence of collecting and retention decisions raises some interesting speculations on the origins behind the relatively high audience level of older materials in library collections. It may be the case that materials of an academic or scholarly nature have the greatest "survivability" in terms of sustaining their perceived value over time, and therefore exhibit a higher retention rate in library collections than materials aimed at a general or juvenile readership, which may be more likely to be weeded out of the collection over time. If so, a direct correlation would exist between the age of a print book and its audience level. Another explanation relates to the perceived institutional mission of a library in regard to preservation. Academic research libraries tend to have a strong sense of obligation to preserve the cultural and scholarly record. This would suggest that older materials would have a higher likelihood of being retained by academic libraries, compared to other types of libraries. Again, this would tend to support a direct correlation between the age of the book and its audience level; in this case, however, the high audience level may have less to do with the actual intellectual content of the book, and more to do with the preservation decisions of libraries that retain it in their collections.

Focus on academic libraries: the "G3"

The analysis in the preceding section examines the aggregate collection of US-published print books in WorldCat, representing the combined print book holdings of libraries of all descriptions: academic, public, special, and so on. Given that many digitization activities, including Google Books, rely heavily on the print book collections of academic libraries, it is useful to take a more focused look at the aggregate collection of print books in WorldCat, emphasizing the holdings of academic libraries. To do this, three large academic research library participants in the Google Books program were selected – one from the East Coast, one from the Midwest, and one from the West Coast. Their print book collections were isolated in WorldCat and then aggregated into one combined collection, with duplicate holdings removed. The result is a reasonably representative illustration of the holdings of academic libraries participating in Google's digitization program, with some rough compensations made for local or regional collecting idiosyncrasies. For the sake of brevity, we will refer to this collection as the Google 3 ("G3") collection.

The G3 collection consists of 9.5 million unique items of all descriptions; of these, about 1.9 million, or 20 percent, are US-published print books. The G3 collection therefore extends over approximately 22 percent of the overall US-published print book collection in WorldCat. Table 9 presents the distribution of the G3 US-published print book collection over the major time periods impacting copyright status.

Table 9: Distribution of G3 US-published print books in WorldCat, by major US copyright periods

Pre-1923:	287,246	15 percent
1923-1963:	375,637	20 percent
Post 1963:	1,224,873	64 percent
Unknown/questionable date:	38,128	2 percent

Comparing these results to those in Table 2 suggests that the G3 collection has a slightly higher proportion of materials in the public domain or potentially in the public domain (35 percent) than that exhibited by the overall aggregate collection of US-published print books in WorldCat (31 percent). Indeed, the G3 collection represents a slightly "older" collection than the WorldCat collection, with half of the print books published after 1974 (compared to 1977 for all of WorldCat); two-thirds published after 1961 (compared to 1964 for all of WorldCat); and three-quarters

published after 1948 (compared to 1951 for all of WorldCat). The fact that the G3 collection is perceptibly older than the overall WorldCat collection correlates with the speculation in the previous section that the relatively high audience level for older materials may be at least partially attributable to a sustained value of scholarly or specialist materials in certain disciplines over time, and that academic libraries may exhibit a greater willingness to retain older materials in their collections in order to fulfill a perceived obligation to contribute toward the long-term preservation of the scholarly record.

As with the overall WorldCat collection, books that are potentially in copyright constitute the bulk of the G3 collection. More specifically, this consists of all print books in the G3 collection published during or after 1923: 1.6 million print books, or 83 percent of the total.

The G3 collection of in-copyright print books are associated with about 820,000 unique authors (individuals) who served some creative role in regard to one or more of the print book manifestations in the collection, and therefore could potentially hold some form of copyright authority over future use of the book. Table 10 reports the authors associated with the most print book manifestations in the G3 collection of potentially in-copyright books.

Table 10: Authors associated with the most potentially in-copyright G3 print book manifestations

Author	Number of Print Book Titles
William Shakespeare	901
Harold Bloom	739
Bruce Rogers	644
William Faulkner	554
Marge Piercy	522
Mark Twain	467
Christopher Morley	430
Jacob Neusner	428
John Steinbeck	381
Douglas C. McMurtrie	372

* Note: Bruce Rogers is a noted typographer; due to cataloging conventions, he is often included as sharing in the responsibility for the creation of a work.

As with the overall WorldCat print book collection, the list is dominated by deceased individuals, most of whom are well-known names in literature and therefore likely to be published and re-published frequently over time. In contrast to the WorldCat collection, however, the list in Table 10 is notable for its absence of children's authors, which is likely explained by the fact that the G3 collection is solely comprised of the holdings of three academic libraries.

Approximately 93 percent, or 1.5 million, of the US-published print books in the G3 collection published during or after 1923 are nonfiction, a proportion similar to that found for the WorldCat aggregate collection. Tables 11 and 12 report the subject and audience level breakdown,

respectively, for the G3 collection of nonfiction print books.

Table 11: Subject breakdown, G3 nonfiction print books

History and Auxiliary Sciences	12 percent
Language, linguistics and literature	11 percent
Health and medicine	9 percent
Business and economics	9 percent
Engineering and technology	7 percent
Philosophy and religion	5 percent
Art & architecture:	5 percent
Sociology	5 percent
Law	4 percent
Education	4 percent
Other	23 percent
Unknown	7 percent

Table 12: Audience level breakdown, G3 nonfiction print books*

Schooler:	1 percent
General:	21 percent
Scholar:	78 percent

* For manifestations where it was possible to calculate an audience level (99 percent of total).

The breakdown of the G3 nonfiction collection by subject suggests a heavier emphasis by academic libraries on collecting titles in History, Language/Linguistics/Literature, and Health/Medicine than in the corresponding WorldCat collection (see Table 4). More striking, however, is the difference in audience level proportions across the "Schooler", "General", and "Scholar" categories. The results in Table 12 suggest that the G3 collection is strongly oriented toward a scholarly or specialist audience: more than three-quarters of the titles fall into this category, compared to a little over half for corresponding WorldCat collection. This is certainly to be expected, given that the G3 collection represents the holdings of three academic libraries, whose primary function is, of course, to serve researchers and learners. However, given that Google Books and other digitization efforts are heavily engaged with academic libraries and their print book collections, it also suggests that current digitization activities may be on a path to produce a resource predominantly of interest to researchers and students, rather than a general readership.

Conclusion

This article characterizes the aggregate collection of US-published print books in WorldCat, with a special emphasis on materials published during or after 1923, and therefore either potentially or definitely in copyright. Findings from the analysis indicate that the collection of US-published print books in WorldCat is quite large, encompassing about 15.5 million print books. Nearly two-thirds of these – those published after 1963 – have a high likelihood of being in copyright; less than 15 percent – those published prior to 1923 – are almost certainly in the public domain, with the rest – those published between 1923 and 1963 – potentially in copyright if copyright was renewed. The post-1923 materials collectively account for more than 80 percent, or about 12.6 million, of the US-published print books in WorldCat. It is difficult to predict how many of these print books might be orphan works, but even a small fraction would, in terms of absolute numbers, be considerable, and require a substantial effort to investigate and clear copyright. One study, based on an examination of a random sample of books, estimates a cost of approximately $200 for each title for which digitization and access permissions were obtained.[14]

Analysis of the post-1923 print books in WorldCat suggests significant limitations to automated assessment of copyright status using bibliographic data. Difficulties arise in operationalizing apparently simple concepts: the simple assertion *"this book was first published in the United States"* can be challenged in terms of the definitions of "book" and "published"; uncertainty can even exist over a book's original country of publication.[15] More generally, assertions that we might like to make about information resources in the context of new issues and questions are not always easily generated from existing data sources built for other purposes. While automated analysis of bibliographic data is useful for establishing the general contours of a large collection of print books in terms of copyright status, it is likely insufficient for making a definitive assessment of any one book's copyright status. Manual intervention will almost certainly be required in many cases, especially if the book turns out to be an orphan work.

Investigations aimed at determining copyright status are becoming more prominent in the procedures and workflows of libraries and other organizations. A recent OCLC Research report found that even as these investigations become more common, much ambiguity still surrounds this work in regard to reliable sources of copyright evidence, procedural due diligence, and benchmarks for decision-making.[16] Often, no single source of information exists to establish an item's "copyright provenance", and institutions invoke different rules and criteria for arriving at a copyright status assessment. At this point, copyright investigations seems to be more *ad hoc* than formulaic, more art than science, and oriented toward minimizing risk rather than achieving certainty. The labor intensity – and by extension, the time and expense – associated with copyright investigations underscores the importance of finding ways to reduce costs: for example by sharing the results of copyright investigations to reduce duplicative effort.

Another important finding from the analysis is the prominence of academic institutions as both suppliers and consumers of mass digitization activities like Google Books. From a supply-side perspective, well over half of the total holdings attached to the 15.5 million US-published print book manifestations in WorldCat belong to academic institutions, indicating that institutions of this kind will necessarily be important sources of the raw materials – print books – needed to supply mass digitization activities. Indeed, most of the current participants in the Google Books library program are academic institutions. From a demand-side perspective, the nature of the materials residing in the collections of academic institutions are, of course, tailored to fit the needs of a research- or specialist-oriented audience, as evidenced by the audience level calculations for the "G3" nonfiction print book collection. Digitization activities operating primarily on the print book holdings of academic institutions will produce digital resources predominantly of interest to academic audiences.

Copyright and regulatory regimes define the limits of what can be done with an information resource. Computing and network technologies afford much greater opportunity to replicate,

distribute, access, and re-purpose information, and as a consequence, views on what these limits should be have been subject to much wider interpretation. Debate over initiatives like the proposed Google book settlement will help shape these limits, but an important element of the discussion is a thorough understanding of the scope and characteristics of the in-copyright materials in library collections.

Acknowledgements

The authors thank our OCLC Research colleague Jenny Toves for the data on unique authors presented in Tables 3 and 10. Thanks also to Peter Hirtle and Michael Cairns for reading an earlier draft of this article, and providing many helpful comments and suggestions.

Notes

1. One of the key components of the Google book settlement is a mechanism for Google to provide for-fee access to digitized copies of in-copyright, out-of-print books. A portion of the fees collected will be allocated to the Book Rights Registry, which would be responsible for distributing them to rights holders. For an overview of the Google book settlement, visit the informational web site <http://www.googlebooksettlement.com/>. For a summary of the settlement's possible implications for libraries, see: Erway, R. (2009) *Impact of the Google Book Settlement on Libraries (Revised Version)*. Report produced by OCLC Research. Published online at: <http://www.oclc.org/programs/publications/reports/2009-01.pdf>.

2. Lavoie, B., L. Connaway, and L. Dempsey (2005) "Anatomy of Aggregate Collections: The Example of Google Print for Libraries" *D-Lib Magazine* Vol. 11, No. 9. Available at: <doi:10.1045/september2005-lavoie>.

3. United States Copyright Office (2006) *Report on Orphan Works*. Washington, DC., p. 15. Available online at: <http://www.copyright.gov/orphan/orphan-report-full.pdf>.

4. Works published outside the US are also covered by US copyright law. However, the rules for determining US copyright status for works published overseas are different than those pertaining to works published in the US. To avoid confusion in the analysis which follows, this study is confined to US-published print books only. This is not to diminish the importance of works published outside the US in digitization activities; indeed, an interesting follow-up study could focus on the scope and characteristics of in-copyright print books published overseas.

5. For consistency, this article uses the same definition of "book" as our earlier article on the "Google 5" (Lavoie, Connaway, Dempsey (2005)). Books are defined as *monographic language materials*. Operationally, in the context of a MARC21 record, a book is identified by the codes "a" and "m" in bytes 6 and 7 of the record leader, respectively. Records describing books in print format were identified by eliminating all non-print formats, such as digital, microform, Braille, and so on. Theses/dissertations and government documents are excluded from the analysis, since these materials are usually acquired and managed as separate segments of the library collection. Theses and dissertations could be an important part of digitization efforts like Google Books and may warrant separate study. US copyright law often treats government documents differently than other publications; for example, a work produced by the federal government is generally considered in the public domain upon publication.

6. The term *manifestation* is formally defined in the FRBR model. See: IFLA Study Group on the Functional Requirements for Bibliographic Records (2009) *Functional Requirements for Bibliographic Records*, p. 21. Available at: <http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf>.

7. See <http://www.diglib.org/forums/spring2009/presentations/HathiTrust.pdf>. HathiTrust also found that of the 2.8 million volumes digitized as of April 2009, about 15 percent were in the public domain.

8. See Ringer, B. (1961) *Study No. 31: Renewal of Copyright*. (Library of Congress: Washington, DC) (reprint).

9. Author names were extracted from the MARC 100 and 700 fields. Figures reported include personal names only; corporate authors are not included.

10. Subject categories are adapted from the OCLC Conspectus divisions. For more information, see: <http://www.oclc.org/support/documentation/collectionanalysis/using/introduction/introduction.htm#conspectus_WCA>.

11. To learn more about the audience level metric, see: <http://www.oclc.org/research/activities/audience/default.htm>. See also: O'Neill, E., L. Connaway, and T. Dickey (2008) "Estimating the Audience Level for Library Resources." *Journal of the American Society for Information Science and Technology*, 59,13. 2042-2050. Pre-print available online at: <http://www.oclc.org/research/publications/archive/2008/oneill-jasist.pdf>.

12. Note that the 2007 cut-off date refers to the time the book's record was entered into WorldCat, not when it was published. It is possible, for example, that a book published in 1995 would not get entered into WorldCat until several years later.

13. The qualifier "high likelihood" is included because materials published before 1989 required a copyright notice to be considered "in copyright". After 1989, the copyright notice requirement was dropped. Therefore, we must assume that some fraction of post-1963 US-published books were published without a copyright notice, and would therefore be in the public domain.

14. Covey, D.T. (2005) *Acquiring Copyright Permission to Digitize and Provide Open Access to Books* (CLIR: Washington, DC) Available at: <http://www.clir.org/pubs/reports/pub134/pub134col.pdf>.

15. Hirtle, P. (2008) "Copyright Renewal, Copyright Restoration, and the Difficulty of Determining Copyright Status" *D-Lib Magazine*, Vol. 14, No. 7/8. Available at: <doi:10.1045/july2008-hirtle>.

16. Proffitt, M., A. Arcolio, and C. Malpas (2008) *Copyright Investigation Summary Report*. Published online at: <http://www.oclc.org/programs/publications/2008-01.pdf>.

Top | Contents
Search | Author Index | Title Index | Back Issues
Editorial | Next Article
Home | E-mail the Editor

---

doi:10.1045/november2009-lavoie